Thomas J. Loredo
Senior Research Associate
Space Sciences Building, Room 620
Ithaca, NY • 14853–6801
607-255-6564; -6918 (fax)
loredo@astro.cornell.edu

March 21, 2008

AISR Program
NASA Headquarters
Washington, DC 20546

To whom it may concern:

This is the final report describing the work of my team on the project titled "Astrostatistical Tools in Python" funded by NASA's AISR program (grant NAG5-12082). This report broadly summarizes our overall software development effort, and more specifically describes activities during the NCE period from Spring 2005 through Spring 2006 (i.e., the period since the last report). Most of the specific accomplishments reported earlier are not duplicated here.

Our project—dubbed the `Inference` project—has as its ovearching goal improving the statistical practice of astronomers, to be accomplished via work in two directions: developing ready-to-use software implementations of both advanced and conventional statistical methods (to facilitate comparison and experimentation); and performing significant outreach to astronomers to educate them about advanced methods. More specifically, the project aimed to accomplish this via a three-pronged program:

- Development of the `Inference` package, a software package for the Python language implementing advanced statistical methods. The package has two main components: (1) A *library*, comprised of multiple modules that implement a variety of statistical algorithms as largely self-contained functions and classes, as well as supporting utility modules implementing computational methods important for statistics. (2) A *Parametric Inference Engine* (PIE), implementing parametric model fitting and assessment in a very flexible object-oriented framework that allows comparison of rival approaches with a unified interface (e.g., chi-squared fitting, maximum likelihood, Bayesian inference).
- Implementation of a multi-faceted outreach program, providing astronomers with statistical training via workshops, special sessions at conferences, and other opportunities.
- Enhancement of the numerical computation capabilities of the Python language by supporting the development of its array and scientific computing packages.

PI Thomas Loredo (Cornell) had lead responsibility for the `Inference` package software development, with Co-I Alanna Connors (Eureka Scientific) sharing responsibility (with planned emphasis on documentation and development of tutorial examples). Connors had lead responsibility for the outreach effort. Finally, Co-I Travis Oliphant (at Brigham Young University during the project; presently at Enthought Scientific Computing Solutions) received modest funding to serve as a Python consultant and to support his work as a lead developer of Python's numerical packages.

The Inference Package

The `Inference` package currently consists of the following code:

- Nearly 10,000 lines of executable Python
- Over 2,500 lines of inline Python documentation strings and 2000 lines of source code comments to assist users
- About 4,500 lines of C code implementing computationally efficient Python extensions
- Over 10,000 lines of Fortran-77 code implementing computationally efficient Python extensions

- About 6,000 lines of "sandbox" Fortran code implementing algorithms awaiting finalized Python interfaces

Nearly all of the Python code and over 80% of the C code is original to our project. About 7,000 lines of the Fortran code comprise 3rd-party tools for which we have written Python interfaces; the remaining Fortran code is original to the project.

The package is hosted at a web site at the following URL, where there is a description of its capabilities and contents:

- `http://inference.astro.cornell.edu/`

The web site itself is automatically generated using a mixture of custom and 3rd-party Python scripts. We are continually adding both software and documentation to the site.

As noted on the web site, the `Inference` package is currently designated as beta-quality code. The code is divided between a main package containing well-tested code that is near release quality, and a smaller "sandbox" containing code whose interfaces are still evolving. The main reason for the "beta" designation is the paucity of web-based documentation and examples (i.e., documentation other than the Python documentation strings available to users at the command line with Python's "help" function, duplicated on the web site as API documentation). Our plan was to have a more extensive collection of web-based documentation and examples. Unfortunately, as noted in previous reports, Connors' contribution to the project was seriously impacted by a cancer remission detected in Fall 2003. Her recovery was prolonged; she first hoped to resume full participation in Fall 2004, but as we reported in the NCE request, this was overly optimistic. Although she was able to resume her time commitment to the project during the NCE period, her effort had to be directed nearly entirely to the outreach program (described below), instead of sharing in code development duties with Loredo in the 2:1 ratio planned in the proposal. This had its biggest impact on the package documentation.

During the NCE period, Connors' code effort focused on developing a Python "front end" for a mixture of C and R code written by David van Dyk and David Esch, implementing the EMC2 Bayesian multiscale Poisson image analysis algorithm developed by van Dyk, Esch, Connors, and collaborators (mentioned in our 2004 report). This algorithm analyzes photon counting image data with rigorous Poisson statistics, and provides error estimates for the restored image. This code has been used to analyze Chandra imaging data. It has been distributed to a few astronomer colleagues on request, but it is not part of the `Inference` package (due to its reliance on R).

The end of our grant support has most definitely not meant the end of our work on the `Inference` package. We continue to add significant capability and documentation to the package, in some cases with support from other programs (when overlap in program goals justifies it), and in some cases without support, simply because of the usefulness of the package and our enthusiasm for its goals. We remain commited to seeing the package mature, particularly in regard to documentation and tutorial examples, and we are eager to remove the self-imposed "beta" designation. As we write, Loredo is adding capability for modeling luminosity and spatial distributions of cosmological sources (to support research on properties of gamma-ray bursts, but designed to support other suvery studies, e.g., of galaxies), and adding MCMC capability to the PIE framework.

Outreach

The NCE period coincided with a unique and important outreach opportunity that deeply engaged both Loredo and Connors. The Statistical and Applied Mathematical Sciences Institute (SAMSI, `www.samsi.info`), an NSF-sponsored institute fostering cross-disciplinary collaboration between statisticians and applied mathematicians, and scientists and engineers, decided to host a semester-long program on astrostatistics in Spring 2006. The Center for Astrostatistics (CASt) at Penn State (also funded by NSF) co-hosted the program.

An overview of the program is at SAMSI's web site:

- `http://www.samsi.info/programs/2005astroprogram.shtml`

Loredo and Connors were invited to help plan and participate in the program, with Connors' participation supported by this grant, and Loredo's supported both by this grant and SAMSI/NSF funds. Loredo became one of the program leaders. The program had four main periods of activity; Loredo and Connors participated in all four:

- A planning session in Summer 2005, held at NASA/Ames and hosted by Jeff Scargle.
- An opening workshop and tutorial series, held at SAMSI in Jan 2006.
- A semester-long research program, centered around the activities of four working groups, taking place both via telecons/web seminars, and intensive sessions at SAMSI.
- The 4th Statistical Challenges in Modern Astronomy (SCMA) workshop, serving as the closing workshop for the program.

The opening workshop was open to the broader astronomical and statistical communities; 67 people attended. Connors and Loredo helped organize the workshop and chaired sessions. The program and talks are archived online:

- `http://astrostatistics.psu.edu/samsi06/index.html`

Loredo organized a 3-day series of tutorials on Bayesian methods for astronomers that followed the workshop; he gave half the lectures, including an introduction to Python and a demonstration of some of the `Inference` package (astronomers William Jefferys and Phil Gregory also lectured). 31 scientists attended these tutorials. Video of the lectures is online at SAMSI:

- `http://www.samsi.info/workshops/2005astro-workshop200601.shtml`

Course notes and the subset of the `Inference` package distributed to participants are archived online at CASt:

- `http://astrostatistics.psu.edu/samsi06/`

Loredo organized two of the four SAMSI working groups: the surveys and population studies (SPS) working group, and the exoplanets (EXO) working group. Loredo led the SPS working group and spent two months as a visiting research associate at SAMSI to support the astrostatistics program. Connors was a key participant in the Source and Feature Detection (SD) working group; the EMC2 algorithm was a focus of attention of this working group. Each group had weekly telecons as well as one- to two-week intensive sessions at SAMSI. The activities of each group, including archived talks and notes, are available online:

- SPS: `http://www.samsi.info/200506/astro/workinggroup/sps/`
- EXO: `http://www.samsi.info/200506/astro/workinggroup/exo/`
- SD: `http://www.samsi.info/200506/astro/workinggroup/sd/`

SCMA IV served as the closing workshop for the program, drawing 104 participants. Though it took place just after the end of our grant's NCE period, both Loredo and Connors helped plan the conference during the grant period and relied on its support. We demonstrated several modules in the `Inference` package during an evening session devoted to software demonstrations. We each also gave talks on astronomical research supported in part by our AISR grant; papers based on these talks were recently published in the ASP conference series (listed below). The conference program and talks are archived at CASt:

- `http://astrostatistics.psu.edu/scma4/program.html`

The SAMSI program was a highlight of our outreach activities and was a fitting culmination for our project. Many astronomers learned about cutting-edge astrostatistics, and we had two forums for demonstrating our software (at SAMSI and SCMA). The working group activity has led to new astronomer/statistician collaborations, including an NSF-funded collaboration between Loredo and SAMSI/Duke statisticians, and a NASA-funded collaboration between Connors and members of the SD working group. Hyunsook Lee, a PSU statistics graduate student (advised by CASt director Jogesh Babu) who participated in several working groups and helped Loredo run the SPS group, was hired as a postdoc at CfA upon graduation; we believe this is the first hire of a PhD statistician by an academic astronomy research group.

Apart from the SAMSI program, we also pursued other outreach opportunities. Connors began co-leading a course in Harvard's Statistics department: *Statistics 310 – Topics in Astrostatistics*. This is an open, interactive, interdisciplinary seminar course, with both invited and contributing speakers from all over. Problems and solutions are proposed and worked on by scientists, students, and statisticians, junior and senior alike. Lectures prepared during the grant period are available online:

- 2004/2005: `http://hea-www.harvard.edu/AstroStat/Stat310_fMMIV/`
- 2005/2006: `http://hea-www.harvard.edu/AstroStat/Stat310_fMMV/`

The students are mostly statisticians wishing to learn more about astronomy, although CfA astronomers (especially from the Chandra collaboration) also attend. The course is still taught, and Connors continues to participate.

Connors also helped organize an astrostatistics session at the Nov 2005 Chandra Calibration workshop, titled, "Incorporating Calibration Uncertainties into Data Analysis." She assisted Vinay Kashyap (CfA), who was the lead organizer.

In June 2005 Loredo began participating in the CASt Summer School in Statistics for Astronomers and Physicists, a one- to two-week long summer school teaching astronomers both basic and advanced statistics. Lecturers include both statisticians and astronomers; the students are predominantly graduate students and postdocs, though senior astronomers also participate. So far, despite minimal advertisement, each year the school is quickly oversubscribed, highlighting the strong demand for this kind of outreach to astronomers. Enrollment is typically three to four dozen students, depending on CASt resources. Loredo has given lectures on Bayesian methods in astronomy at every session; his participation in 2005 and 2006 was supported in part by this grant. The 2005 lectures are archived at CASt:

- `http://astrostatistics.psu.edu/su05/astrostat_courses.html`

Python's `Numerical` Capability

Finally, there was dramatic progress in Python's numerical capability during the NCE period, spearheaded by Co-I Oliphant. This period saw the first release of `numpy`, a new array package for Python, designed and largely developed by Oliphant. This package synthesizes the best features of the earlier `Numeric` package (originally developed at LLNL and MIT), and the recent `numarray` package developed at STScI. `Numeric` performed well with small arrays that could fit in RAM, but could not support large, memory-mapped arrays needed by astronomers for analyzing large images and spectra. The `numarray` package provided solid support for large arrays, at the cost of reduced performance with small arrays. Many scientists were unwilling to suffer the reduced performance and would not move to `numarray`, leading to a split in the Python scientific computing community. Oliphant took the best ideas from `Numeric` and `numarray` and produced a package better than either. It was first released in Winter 2005/2006; Loredo quickly adapted the `Inference` package to `numpy` and his Python demo at SAMSI included one of the earliest presentations of `numpy`. In the time since, `Numeric` and `numarray` have been deprecated (even by STScI) and the Python scientific computing community is now unified in its array package. Oliphant also made significant improvements to SciPy, a package that builds on `numpy`, providing tools essential to scientific computing such as optimizers, special functions, random number generators, etc.. The `Inference` package relies heavily on both `numpy` and SciPy.

Oliphant also proved invaluable as a consultant. The grant's support of Oliphant was modest (1/12 FTE/yr), so we cannot take too much credit for his work, but we are grateful that AISR could play a role in supporting such a fundamental enhancement of Python. It has had an significant impact on the entire broad and growing community of scientists and engineers who rely on Python for numerical computation.

Recent Publications

Recent and in-press publications documenting work supported in part by this grant, written since the last report, include three papers describing work of the SD, EXO, and SPS SAMSI working groups (published in the SCMA proceedings), a paper using some of the `Inference` package to analyze Kuiper belt object survey data, and three articles (including two invited book chapters) addressing survey data analysis issues in astronomy, respectively:

Connors, A.; van Dyk, D. (2007) "How To Win With Non-Gaussian Data: Poisson Goodness-of-Fit" Statistical Challenges in Modern Astronomy IV, ASP Conference Series, Vol. 371, 101

Clyde, M. A.; Berger, J. O.; Bullard, F.; Ford, E. B.; Jefferys, W. H.; Luo, R.; Paulo, R.; Loredo, T. J. (2007) "Current Challenges in Bayesian Model Choice" Statistical Challenges in Modern Astronomy IV, ASP Conference Series, Vol. 371, 224

Loredo, T. J. (2007) "Analyzing Data from Astronomical Surveys: Issues and Directions" Statistical Challenges in Modern Astronomy IV, ASP Conference Series, Vol. 371, 121

Petit, J.-M.; Holman, M. J.; Gladman, B. J.; Kavelaars, J. J.; Scholl, H.; Loredo, T. J. (2006) "The Kuiper Belt luminosity function from mR= 22 to 25" MNRAS, 365, 429-438

Loredo, T. J. (2004) "Accounting for Source Uncertainties in Analyses of Astronomical Survey Data" in 24th International Workshop on Bayesian `Inference` and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings, Vol. 735, 195–206

Petit, J.-M., Kavelaars, J. J., Gladman, B., Loredo, T. (2008) "Size Distribution of Multikilometer Transneptunian Objects" chapter in The Solar System Beyond Neptune (ed. M. A¿ Barucci, H. Boehnhardt, D. P. Cruikshank, A. Morbidelli), U. of Arizona Press, in press (17pp)

Loredo, T. J., and Hendry, M. A. (2008) "Bayesian Multilevel Modelling of Cosmological Populations" chapter in Bayesian Methods in Cosmology (ed. A. Liddle et al.), Cambridge University Press, in press (18pp)

This grant supported the early stages of some of our work in progress which will soon produce publications that will be citing this grant's prior support.

As always, we are very grateful to the AISR program for its support of our work.

Thank you,

Thomas J. Loredo